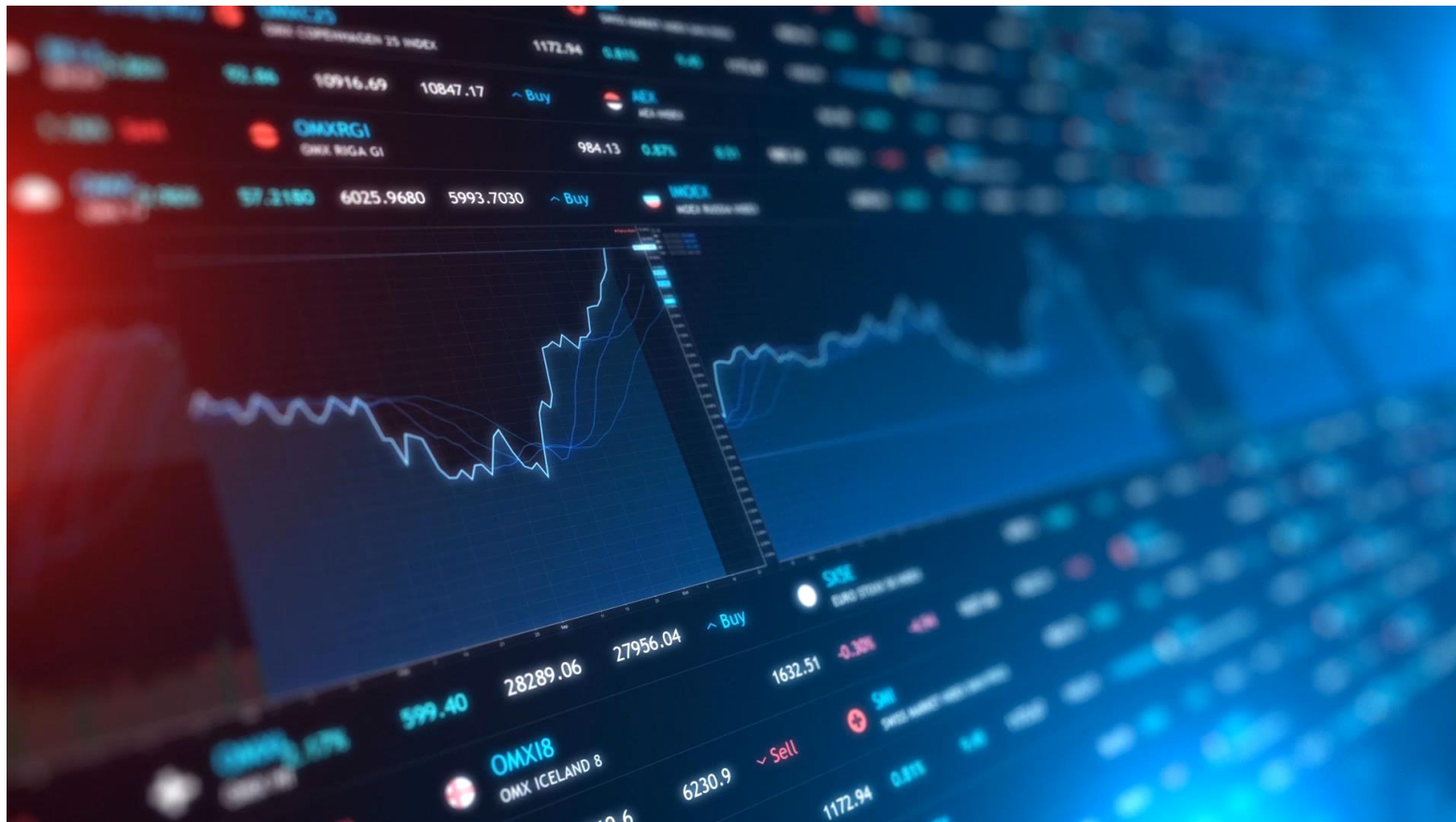


Wine Analysis



Team Members

1)Bharath Kumar Uppala

2) Prudhvish Narayanam

3)V.Yagnesh

4)LV Tharun Kumar Reddy





Predicting Wine Quality with Several Classification Techniques

Table of Content

1. [Introduction](#)
2. [Setup](#)
3. [Exploring Variables](#)
4. [Convert to a Classification Problem](#)
5. [Preparing Data for Modelling](#)
6. [Modelling](#)
7. [Feature Importance](#)



Introduction

For this project, I used Kaggle's [Red Wine Quality](#) dataset to build various classification models to predict whether a particular red wine is "good quality" or not.

Each wine in this dataset is given a "quality" score between 0 and 10. For the purpose of this project, I converted the output to a binary output where each wine is either "good quality" (a score of 7 or higher) or not (a score below 7). The quality of a wine is determined by 11 input variables:

Contents of wine

Fixed
acidity

Volatile
acidity

Citric acid

Residual
sugar

Chlorides

Free sulfur
dioxide

Total sulfur
dioxide

Density

pH

Sulfates

Alcohol

Objectives

The objectives of this project are as follows



```
graph TD; A[The objectives of this project are as follows] --> B[To experiment with different classification methods to see which yields the highest accuracy]; B --> C[To determine which features are the most indicative of a good quality wine];
```

To experiment with different classification methods to see which yields the highest accuracy

To determine which features are the most indicative of a good quality wine

Setup

- First, I imported all of the relevant libraries that I'll be using as well as the data itself.
- Importing Libraries

```
import numpy as np
import pandas as pd
import matplotlib as plt
import seaborn as sns
import plotly.express as px
```


- Reading Data

```
df = pd.read_csv("../input/red-wine-quality-cortez-et-al-2009/winequality-red.csv")
```

- Understanding Data

```
# See the number of rows and columns
print("Rows, columns: " + str(df.shape))

# See the first five rows of the dataset
df.head()
```

```
Rows, columns: (1599, 12)
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Missing Values

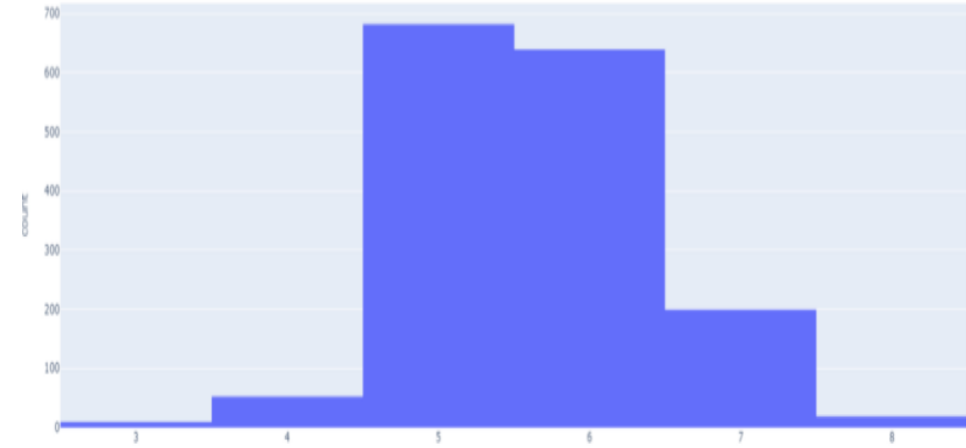
This is a very beginner-friendly dataset. I did not have to deal with any missing values, and there isn't much flexibility to conduct some feature engineering given these variables. Next, I wanted to explore my data a little bit more.

```
fixed acidity      0  
volatile acidity  0  
citric acid       0  
residual sugar    0  
chlorides         0  
free sulfur dioxide 0  
total sulfur dioxide 0  
density          0  
pH               0  
sulphates        0  
alcohol          0  
quality          0  
dtype: int64  
There are no missing values
```

```
# Missing Values  
print(df.isna().sum())
```

Exploring Variables

- **Histogram of 'quality' variable**
- First, I wanted to see the distribution of the *quality* variable. I wanted to make sure that I had enough 'good quality' wines in my dataset — you'll see later how I defined 'good quality'.

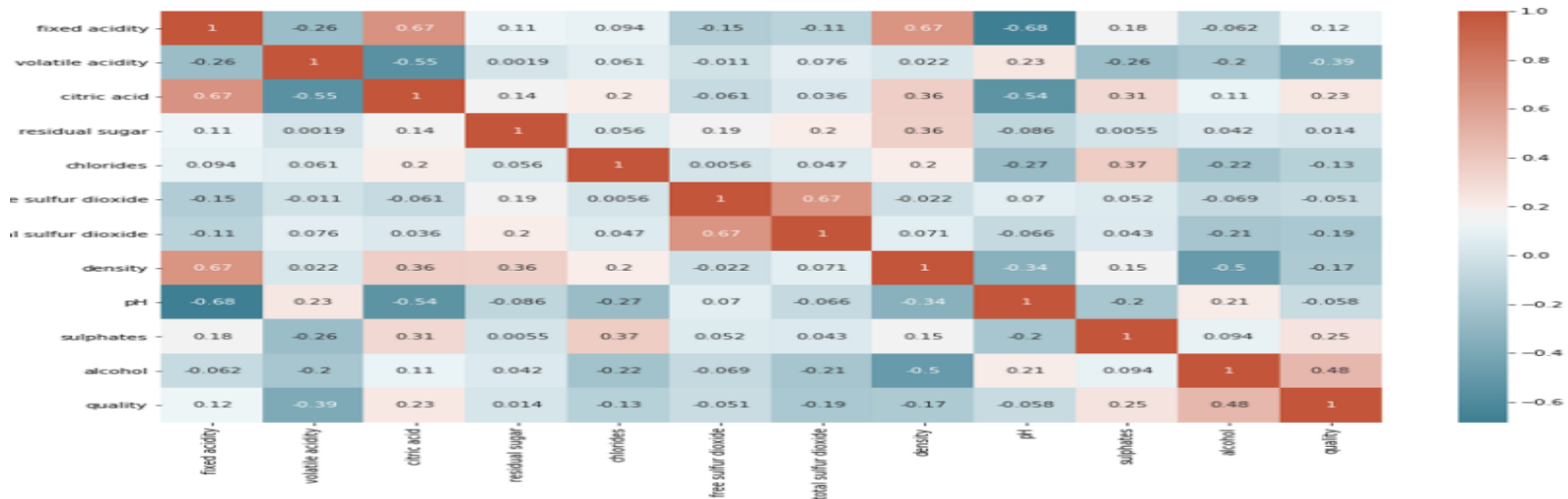


```
fig = px.histogram(df, x='quality')  
fig.show()
```

Correlation Matrix

- Next I wanted to see the correlations between the variables that I'm working with. This allows me to get a much better understanding of the relationships between my variables in a quick glimpse.
- Immediately, I can see that there are some variables that are strongly correlated to *quality*.

```
corr = df.corr()  
matplotlib.pyplot.subplots(figsize=(15,10))  
sns.heatmap(corr, xticklabels=corr.columns, yticklabels=corr.columns,  
            annot=True, cmap=sns.diverging_palette(220, 20, as_cmap=True))
```



Convert to a Classification Problem

- For this problem, I defined a bottle of wine as 'good quality' if it had a quality score of 7 or higher, and if it had a score of less than 7, it was deemed 'bad quality'.
- Once I converted the output variable to a binary output, I separated my feature variables (X) and the target variable (y) into separate dataframes.

```
# Create Classification version of target variable
df['goodquality'] = [1 if x >= 7 else 0 for x in df['quality']]

# Separate feature variables and target variable
X = df.drop(['quality', 'goodquality'], axis = 1)
y = df['goodquality']
```

Proportion of Good vs Bad Wines

- I wanted to make sure that there was a reasonable number of good quality wines. Based on the results below, it seemed like a fair enough number. In some applications, resampling may be required if the data was extremely imbalanced, but I assumed that it was okay for this purpose.

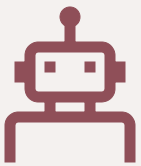
```
# See proportion of good vs bad wines  
df['goodquality'].value_counts()
```

```
0    1382  
1     217  
Name: goodquality, dtype: int64
```

Preparing Data for Modelling

- **Standardizing** the data means that it will transform the data so that its distribution will have a mean of 0 and a standard deviation of 1. It's important to standardize your data in order to equalize the range of the data.

```
# Splitting the data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test size=.25, random state=0)
```

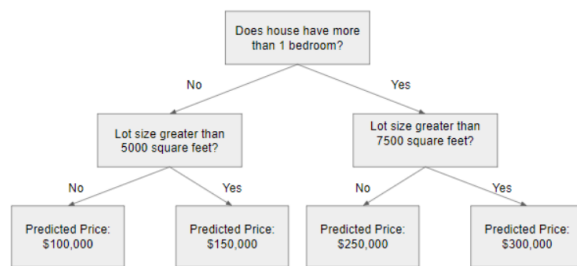



Modelling

- For this project, I wanted to compare five different machine learning models: decision trees, random forests, AdaBoost, Gradient Boost, and XGBoost. For the purpose of this project, I wanted to compare these models by their accuracy.

Model 1: Decision Tree

- Decision trees are a popular model, used in operations research, strategic planning, and machine learning. Each square above is called a node, and the more nodes you have, the more accurate your decision tree will be (generally). The last nodes of the decision tree, where a decision is made, are called the leaves of the tree. Decision trees are intuitive and easy to build but fall short when it comes to accuracy.



	precision	recall	f1-score	support
0	0.96	0.92	0.94	355
1	0.53	0.73	0.62	45
accuracy			0.90	400
macro avg	0.75	0.83	0.78	400
weighted avg	0.92	0.90	0.90	400

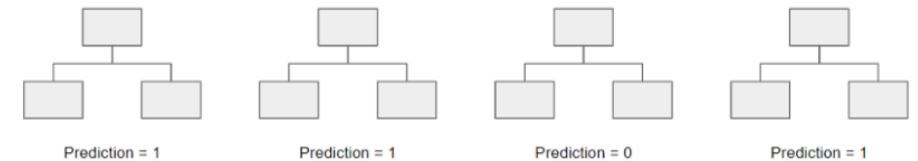
```
from sklearn.metrics import classification_report
from sklearn.tree import DecisionTreeClassifier

modell = DecisionTreeClassifier(random_state=1)
modell.fit(X_train, y_train)
y_pred1 = modell.predict(X_test)

print(classification_report(y_test, y_pred1))
```

Model 2: Random Forest

- Random forests are an [ensemble learning](#) technique that builds off of decision trees. Random forests involve creating multiple decision trees using [bootstrapped datasets](#) of the original data and randomly selecting a subset of variables at each step of the decision tree. The model then selects the mode of all of the predictions of each decision tree. What's the point of this? By relying on a "majority wins" model, it reduces the risk of error from an individual tree.



```
from sklearn.ensemble import RandomForestClassifier
model2 = RandomForestClassifier(random_state=1)
model2.fit(X_train, y_train)
y_pred2 = model2.predict(X_test)

print(classification_report(y_test, y_pred2))
```

Model 3: AdaBoost

- The next three models are boosting algorithms that take weak learners and turn them into strong ones. I don't want to get sidetracked and explain the differences between the three because it's quite complicated and intricate. That being said, I'll leave some resources where you can learn about AdaBoost, Gradient Boosting, and XGBoosting.

	precision	recall	f1-score	support
0	0.94	0.94	0.94	355
1	0.51	0.49	0.50	45
accuracy			0.89	400
macro avg	0.72	0.71	0.72	400
weighted avg	0.89	0.89	0.89	400

```
from sklearn.ensemble import AdaBoostClassifier
model3 = AdaBoostClassifier(random_state=1)
model3.fit(X_train, y_train)
y_pred3 = model3.predict(X_test)

print(classification_report(y_test, y_pred3))
```

Model 4: Gradient Boosting

```
from sklearn.ensemble import GradientBoostingClassifier
model4 = GradientBoostingClassifier(random_state=1)
model4.fit(X_train, y_train)
y_pred4 = model4.predict(X_test)

print(classification_report(y_test, y_pred4))
```

	precision	recall	f1-score	support
0	0.94	0.94	0.94	355
1	0.52	0.51	0.52	45
accuracy			0.89	400
macro avg	0.73	0.73	0.73	400
weighted avg	0.89	0.89	0.89	400

Model 5: XGBoost

```
import xgboost as xgb
model5 = xgb.XGBClassifier(random_state=1)
model5.fit(X_train, y_train)
y_pred5 = model5.predict(X_test)

print(classification_report(y_test, y_pred5))
```

	precision	recall	f1-score	support
0	0.96	0.95	0.95	355
1	0.62	0.69	0.65	45
accuracy			0.92	400
macro avg	0.79	0.82	0.80	400
weighted avg	0.92	0.92	0.92	400



The 4 Factors and 4 Indicators of Wine Quality

Understanding Wine Quality

Wine quality refers to the factors that go into producing a wine, as well as the indicators or characteristics that tell you if the wine is of high quality.

When you know what influences and signifies wine quality, you'll be in a better position to make good purchases. You'll also begin to recognize your preferences and how your favorite wines can change with each harvest. Your appreciation for wines will deepen once you're familiar with wine quality levels and how wines vary in taste from region to region.

Some wines are higher-quality than others due to the factors described below. From climate to viticulture to winemaking, a myriad of factors make some wines exceptional and others run-of-the-mill.

An aerial photograph of a vineyard on a hillside. The rows of grapevines are lush green and follow the contours of the slope. The ground between the rows is covered in dry, brownish vegetation. In the background, a dense forest of tall, dark green trees covers the upper part of the hill. The text "Four Factors That Contribute to Wine Quality" is overlaid in the center in a blue, cursive font.

Four Factors That Contribute to Wine Quality



1. Climate and Weather

- The terroir of wine has a clear-cut influence on its quality. Climate and weather help determine how quickly wine grapes grow, how much flavor and juiciness they have, and how well those grapes can be turned into wine.
- Climate is (relatively) stable, so it's easier for producers to anticipate how climate influences grapes. Cooler climates produce wines higher in acidity but lower in sugar and alcohol. Hotter climates encourage ripening, leading to wines with higher sugars, higher alcohol and fuller body. Producers trying to grow varieties that don't do well in that specific climate will produce a wine of lower quality.
- Weather, on the other hand, can have a more direct, immediate effect on wine quality. It can even spell the difference between a good vintage and a bad vintage. Wines with higher quality will come from grapes that received exactly the inputs they needed.

2. Temperature and Sunlight

To carry out photosynthesis, grape vines must be exposed to temperatures between 60 and 70 degrees Fahrenheit.

In regions with average temperatures closer to 60 degrees F, short-cycle varieties will be successful but long-cycle ones won't. If temperatures are too hot, though, the grapes will ripen too quickly – cutting short the time needed for flavor, color, and other compounds to fully develop.



3. Growing Practices

In addition to what the land and sky provide, the ways in which a producer manipulates the vines will also influence the quality of the resultant wine. Canopy management often includes removing extra leaves and shoots to increase sunlight exposure, while pruning removes select branches to control yields and keep vines healthy.


Harvesting is another crucial factor, since a harvest that is too early or late can lead to grapes lacking their ideal balance. Whether producers harvest manually or mechanically also influences grape quality.

Mechanical harvesting can't apply selective-picking methods, but it does allow for speedy harvests when bad weather threatens to ruin the grapes. Conversely, manual harvesting is slower, but it ensures that only high-quality grapes make it to the winery.



4. Winemaking Practices

- The winemaking process is equally important in determining the final quality of the wine. Wineries follow four main steps when producing their wines, maceration, fermentation, extraction and aging, and they must ensure consistency to get the most from their grapes.
- Inputs such as sulfur dioxide and processing enzymes, as well as decisions with oak barrel aging and oxygen management, all contribute to the quality of wine – from the exceptional to the insipid.

A still life composition featuring wine bottles, glasses, and grapes on a dark wooden surface. In the top left, a bunch of green grapes sits next to a glass of red wine. Below it, a wine bottle with a red foil-wrapped neck lies horizontally. To its right, another wine bottle with a green foil-wrapped neck is partially visible. In the bottom left, a small cluster of green grapes is next to a folded white cloth. The background is a dark, textured wooden surface. A white rectangular box with a dotted border is centered on the right side, containing the title text.

Four Indicators of Wine Quality



1. Complexity

- Higher quality wines are more complex in their flavor profile. They often have numerous layers that release flavors over time. Lower quality wines lack this complexity, having just one or two main notes that may or may not linger.
- With high-quality wines, these flavors may appear on the palate one after the other, giving you time to savor each one before the next appears.



2. Balance

- Wines that have good balance will be of higher quality than ones where one component stands out above the rest.
- The five components – acidity, tannins, sugar/sweetness, alcohol and fruit – need to be balanced. For wines that need several years of aging to reach maturity, this gives them the time they need to reach optimal balance.
- Higher quality wines don't necessarily need moderation in each component – indeed, some red wines have higher acidity while others have a higher alcohol content. What makes the difference is that the other components balance things out.

-

3. Typicity

Another indicator of wine quality comes from typicity, or how much the wine looks and tastes the way it should.

For example, red Burgundy should have a certain appearance and taste, and it's this combination that wine connoisseurs look for with each new vintage. An Australian Shiraz will also have a certain typicity, as will a Barolo, a Rioja or a Napa Valley Cabernet Sauvignon, among others.

4. Intensity and Finish

- The final indicators of both white and red wine quality are the intensity and finish. High-quality wines will express intense flavors and a lingering finish, with flavors lasting after you've swallowed the wine. Flavors that disappear immediately can indicate that your wine is of moderate quality at best. The better the wine, the longer the flavor finish will last on your palate.



Thanks all